

ModSim Challenges for In Situ Workflows

“Workflow: sequencing and orchestrating operations, along with moving data among those operations.” – Deelman et al. 2014.

“Data movement, rather than computational processing, will be the constrained resource at exascale.” – Dongarra et al. 2011.

Context

Scientific Data Analysis Today: Three Assumptions

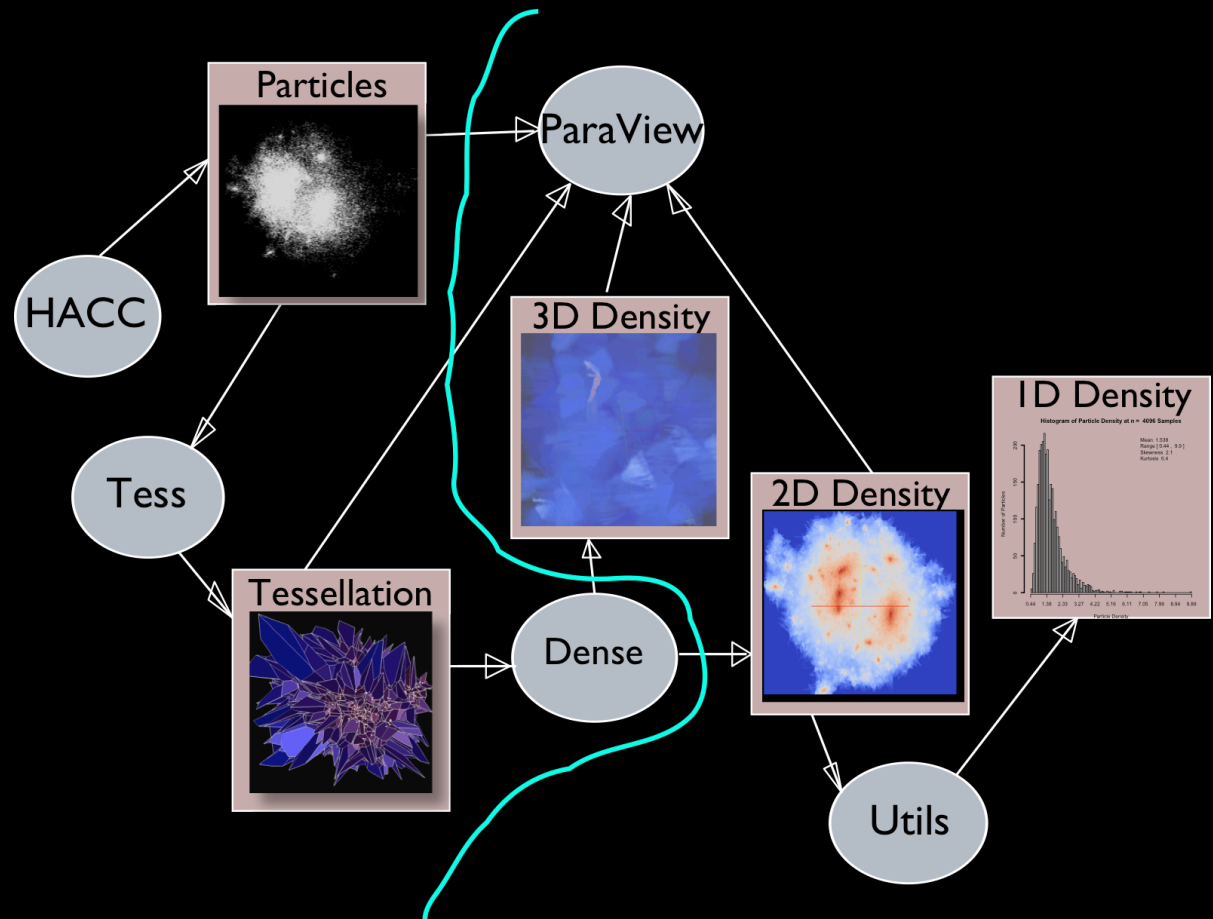
1. Computational science is a complex combination of multiple tasks, i.e., a workflow.
2. Scientific data analysis requires big science resources, i.e., HPC w/ some in situ analysis.
3. HPC is parallel:
 - In one task => data parallelism
 - **Between tasks => task parallelism**

Simple In Situ Workflow Example

Analysis of Cosmology Simulations



- Just one small part of the complete cosmology workflow
- Converts dark matter particles to an unstructured mesh
- Converts an unstructured mesh to a regular grid
- Computes statistics over the grid and visualizes the results



Companion Efforts

April 2015 NGNS/CS
Workshop on the Future of
Scientific Workflows

ENTIRE WEEK SUNDAY MONDAY TUESDAY WEDNESDAY THU

CHARACTERIZING EXTREME-SCALE COMPUTATIONAL AND DATA-INTENSIVE WORKFLOWS

SESSION: Characterizing Extreme-Scale Computational and Data-Intensive Workflows

EVENT TYPE: Birds of a Feather

EVENT TAG(S): Workflows

TIME: 3:30PM - 5:00PM

SESSION LEADER(S): Edward Seidel, Franck Cappello, Tom Peterka

ROOM: 13A

THE FUTURE OF SCIENTIFIC WORKFLOWS

REPORT OF THE DOE NGNS/CS SCIENTIFIC WORKFLOWS WORKSHOP
APRIL 20-21, 2015

Sponsored by the Office of Advanced Scientific Computing Research
U.S. Department of Energy
Office of Science

U.S. DEPARTMENT OF **ENERGY** Office of Science

WORKS 2016 SC Workshop
11th annual workshop on
workflows in support of
large-scale science

Works 2016 Details Organization Venue Program Links

WORKS 2016 Workshop

Workflows in Support of Large-Scale Science
Monday, 14 November 2016, Salt Lake City, Utah.

Held in conjunction with **SC16** The International Conference for High Performance Computing, Networking, Storage and Analysis

Co-chaired by **Sandra Gesing** University of Notre Dame, USA and **Rizos Sakellariou** University of Manchester, UK.

ISAV 2016 SC Workshop
2nd Annual workshop on in
situ infrastructures for
analysis and visualization

SC15 sig hpc
Austin, TX | hpc transforms.

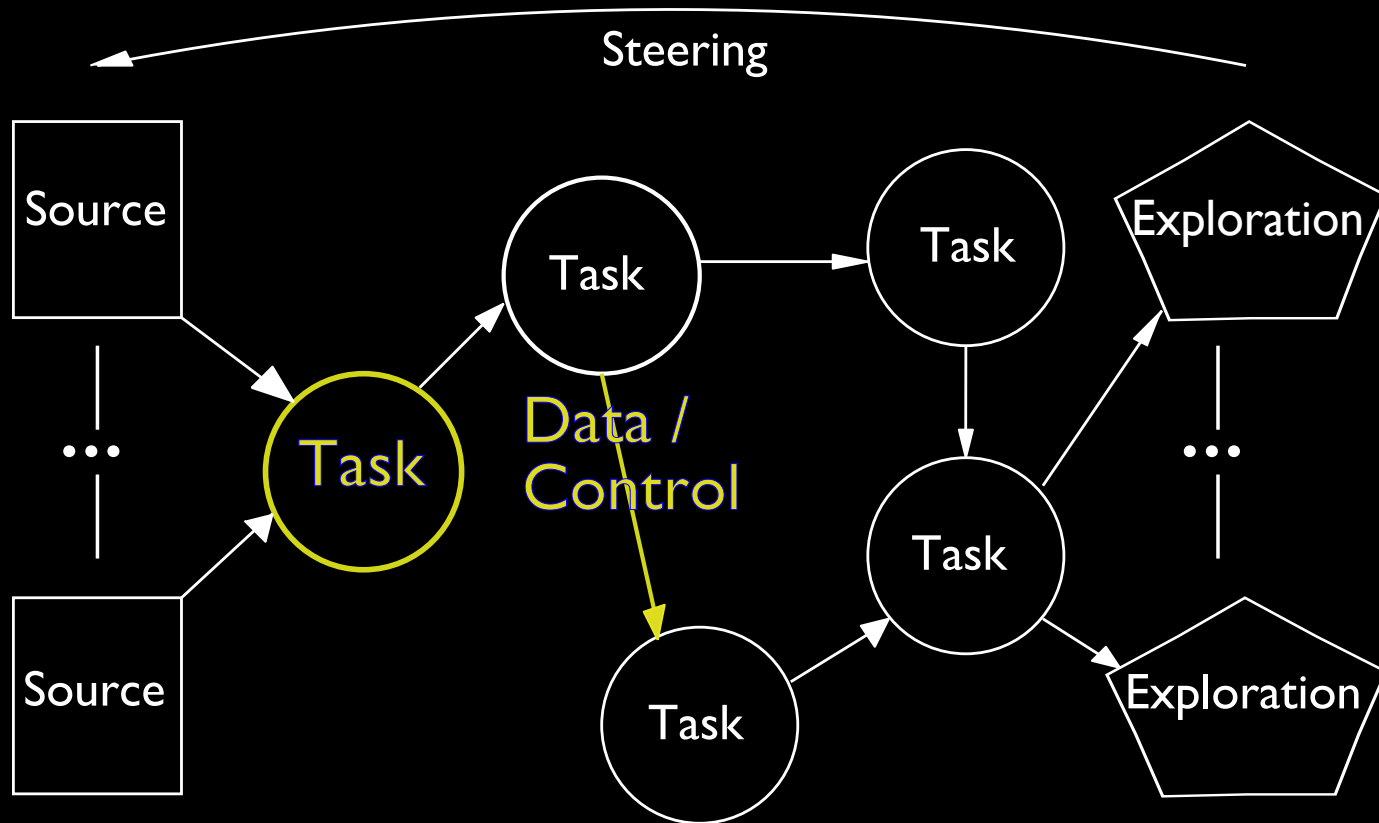
ISAV 2015: In Situ Infrastructures for Enabling Extreme-scale Analysis and Visualization

Terminology

Definitions

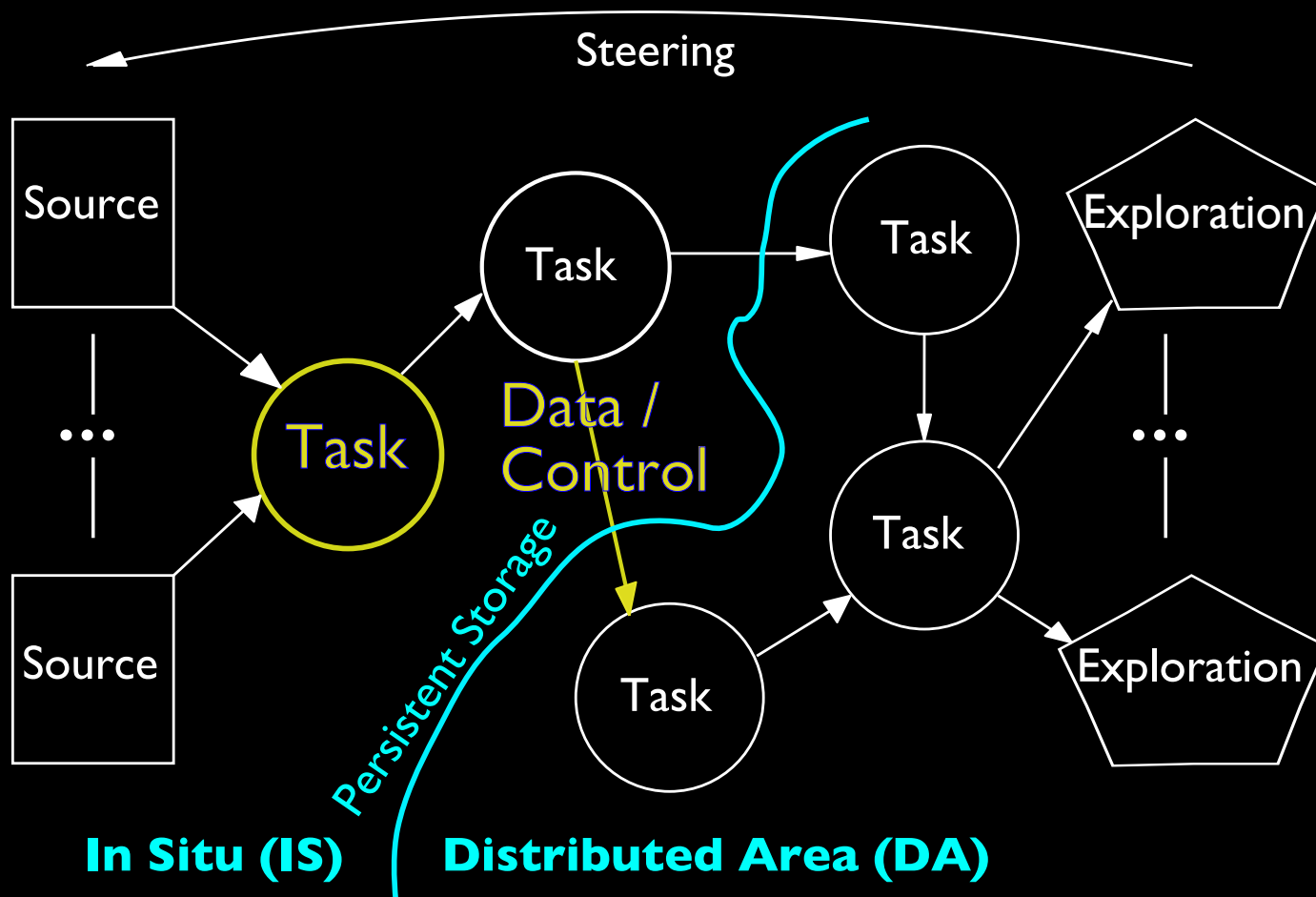
- **Workflow:** Sequencing and orchestrating operations, along with the attendant tasks of, for example, moving data between workflow processing stages. (“programming in the large”)
- **Workflow management systems:** Aiding in the automation and capture the provenance of these processes, freeing the scientist from the details of the process.
 - Manage the execution of constituent tasks
 - Manage the information exchanged between them
- **Usability:** Benefitting the target audience (computational scientists) on target platforms (computing environments) and reused across sciences and computing environments and whose performance and correctness can be modeled and verified.
- **Performance:** Overhead (memory, time, power) compared with tight coupling of inline functions, impact on simulation compared with no in situ analysis, performance compared with models and predictions.
- **Validation:** Accuracy and scientific reproducibility

Workflow = Directed Graph



- **Digraph:** not DAG; i.e., can have cycles
- **Nodes:** tasks (can be parallel)
- **Links:** communication between tasks (can be parallel)

In Situ (IS) and Distributed Area (DA)



- **In Situ (IS)**: Within an HPC system (synonyms: in situ, in transit, coprocessing, run-time, online)
- **Distributed Area (DA)**: Across systems, potentially geographically distributed
- **IS** \approx **HPC** (high-performance computing), **DA** \approx **DAIC** (distributed-area instruments and computing)

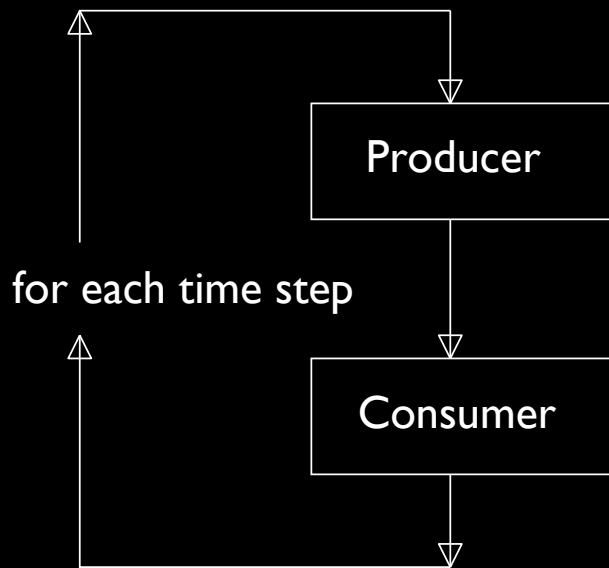
Different Flavors of In Situ

Producer
E.g., simulation



Consumer
E.g., analysis

Conceptual

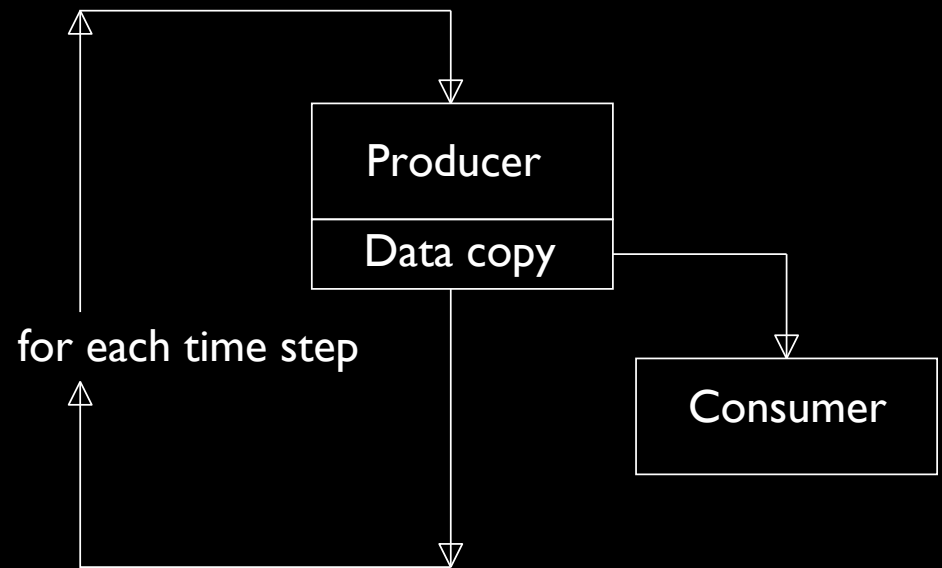


Resources:
A = {Nodes, cores, threads}



Resources: A

Time division



Resources: A



Resources: B

Space division

Data Rates Matter

Producer
E.g., simulation

P

Fast

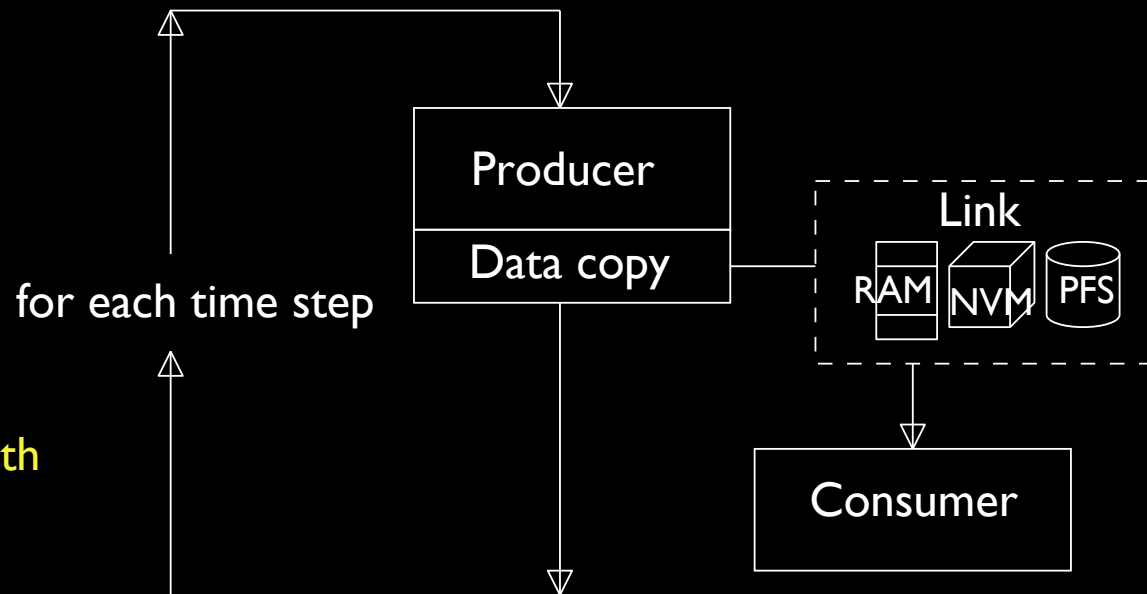
Slow

C

Consumer
E.g., analysis

Conceptual:

Fast producer and slow consumer



Space division with
link resources

Resources: A

P

Resources: B

L

Resources: D

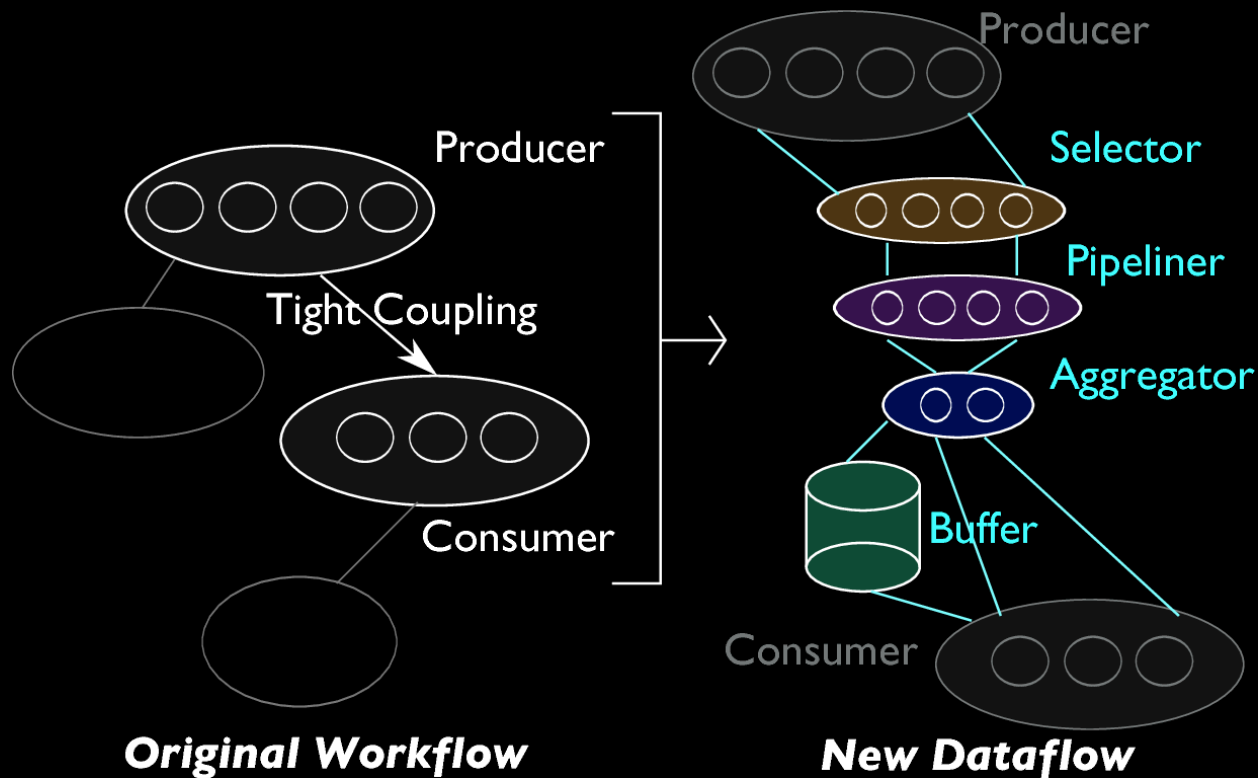
C

Graph representation

From Workflows to Dataflows

Data Movement Between Components

Decoupling by converting a single link into a dataflow enables new features such as fault tolerance and improved performance.



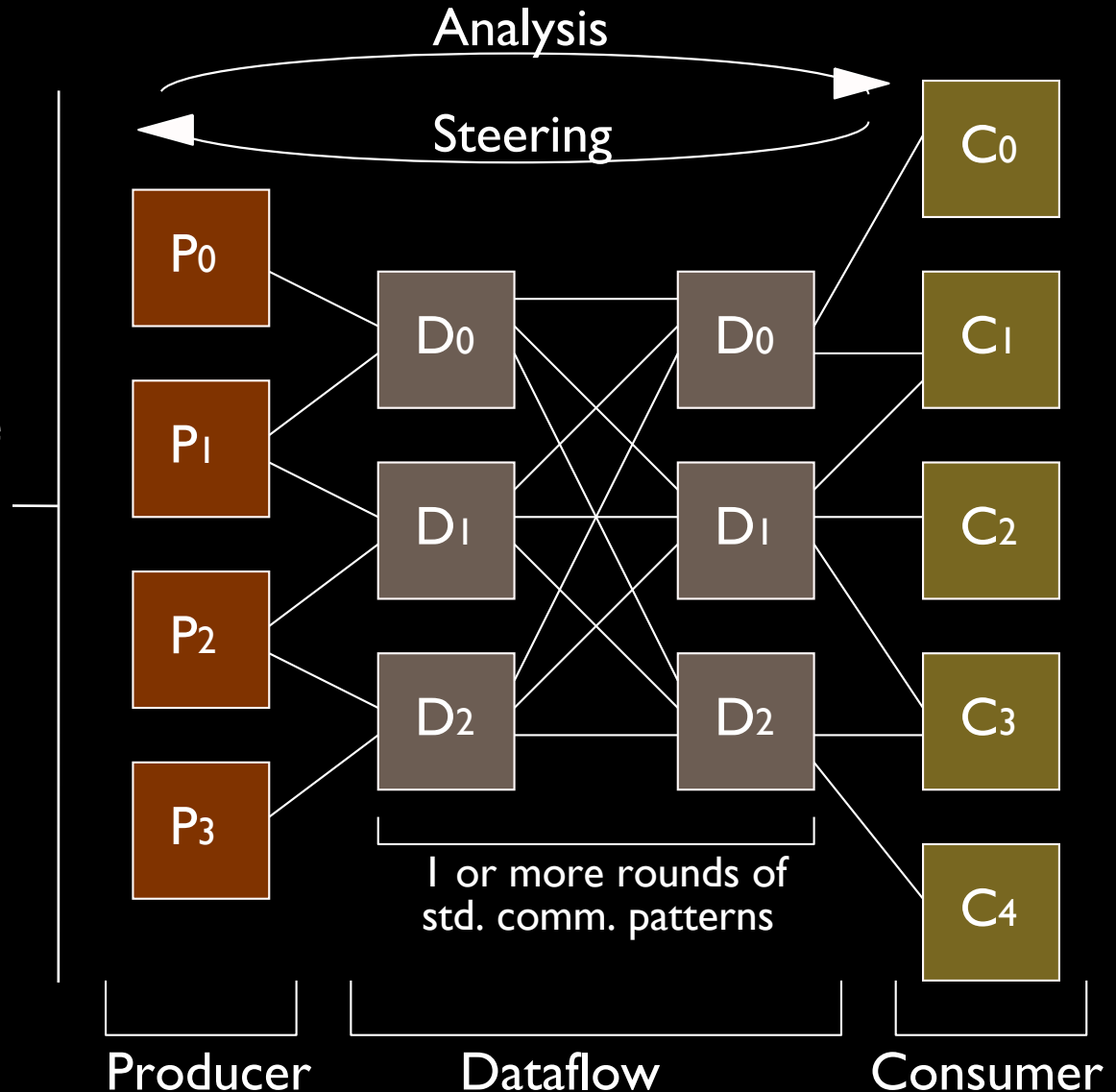
Links and Dataflows

The link can be a simple noop or a complete parallel program performing complex data transformations.

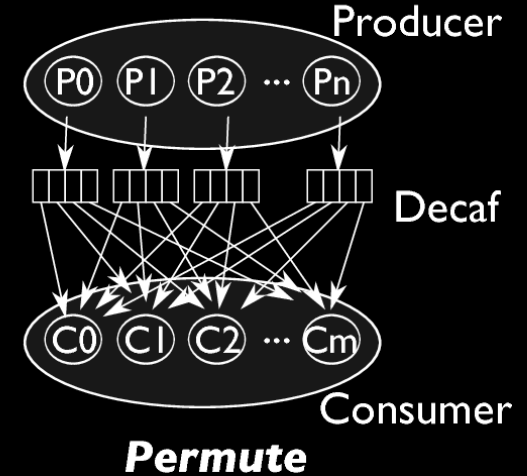
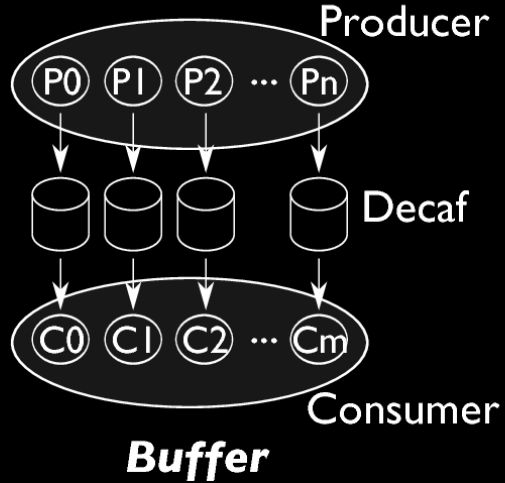
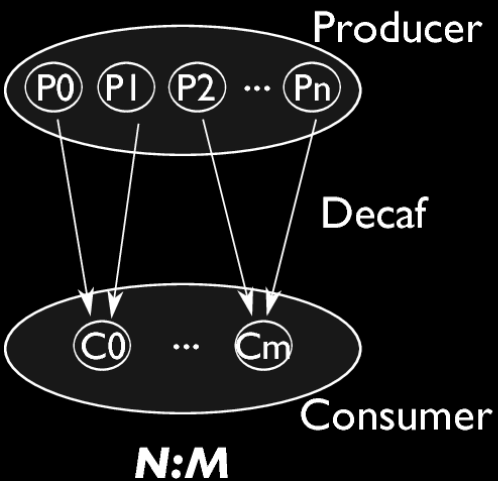
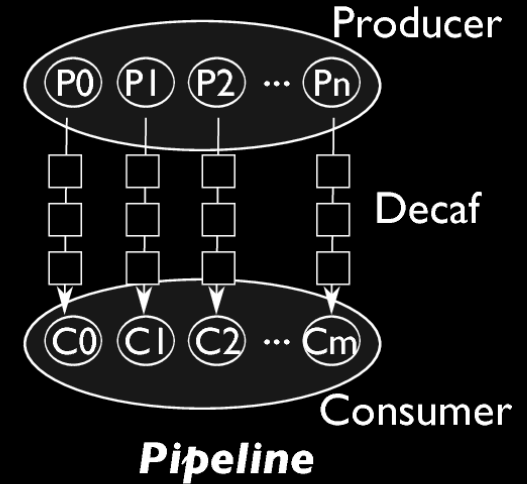
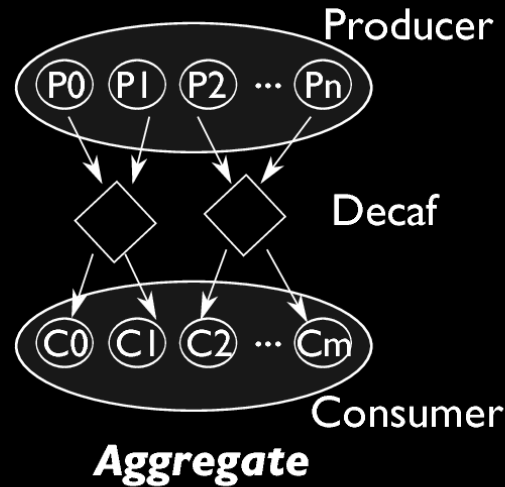
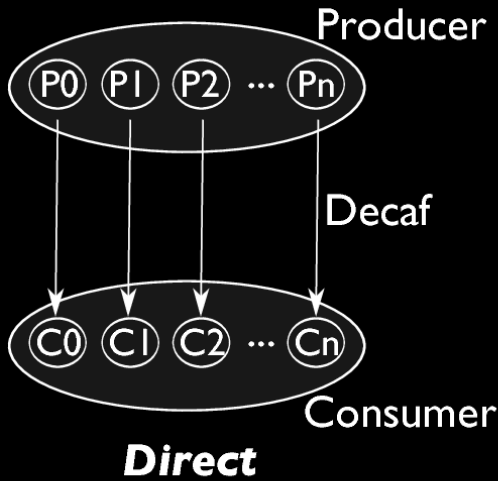
The link can be part of the producer, consumer, or have its own resources.

Dataflow across producer, link, consumer is parallel.

Producer, link, and consumer are also parallel programs with their own internal communication.

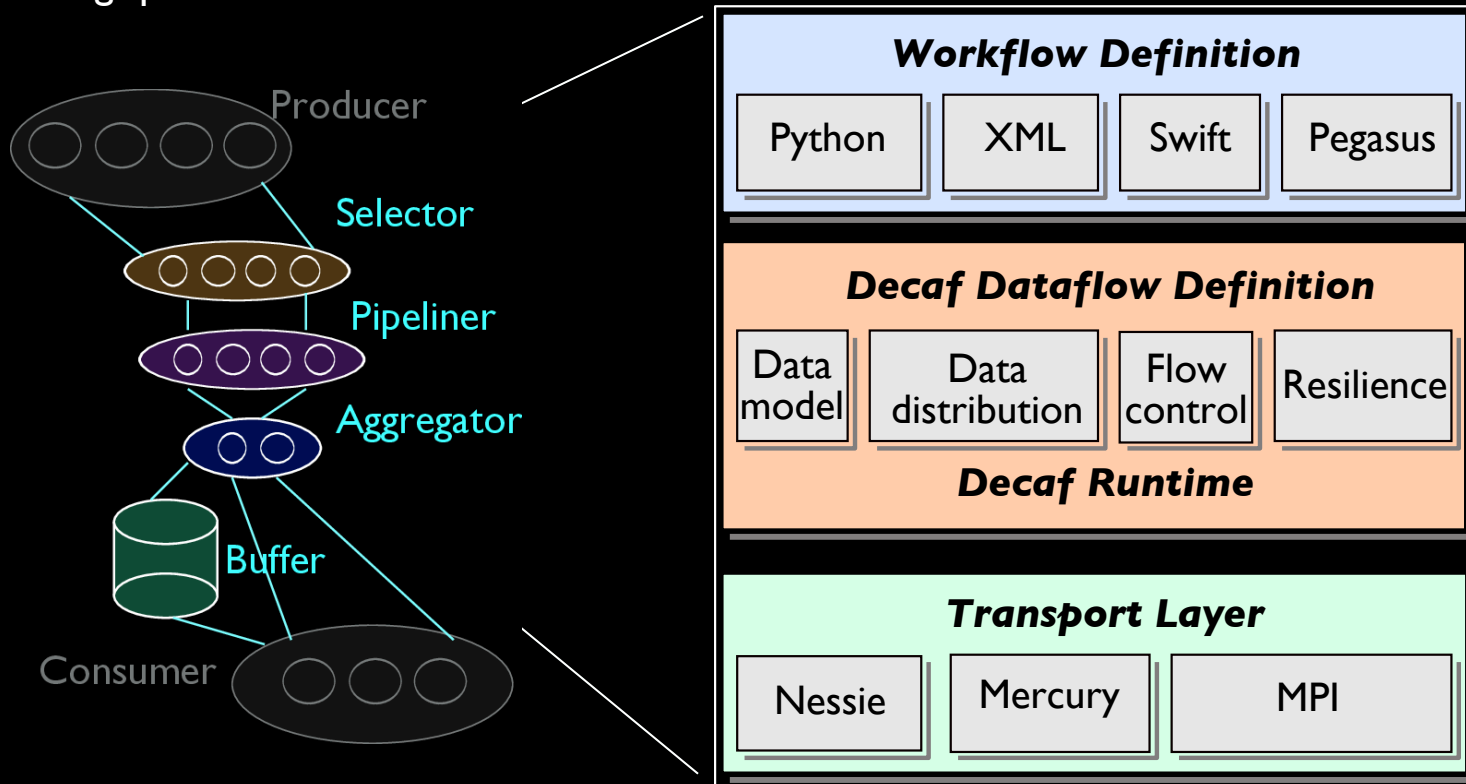


Dataflow Design Patterns



Dataflow modes include aggregation, pipelining, and automatic buffering while potentially permuting data in an N:M and direct coupling of parallel codes.

Decaf: Decoupled Dataflows

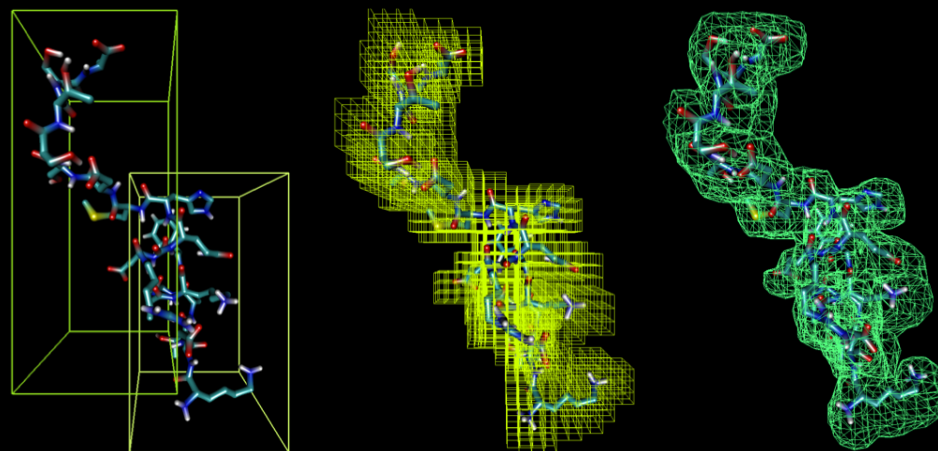
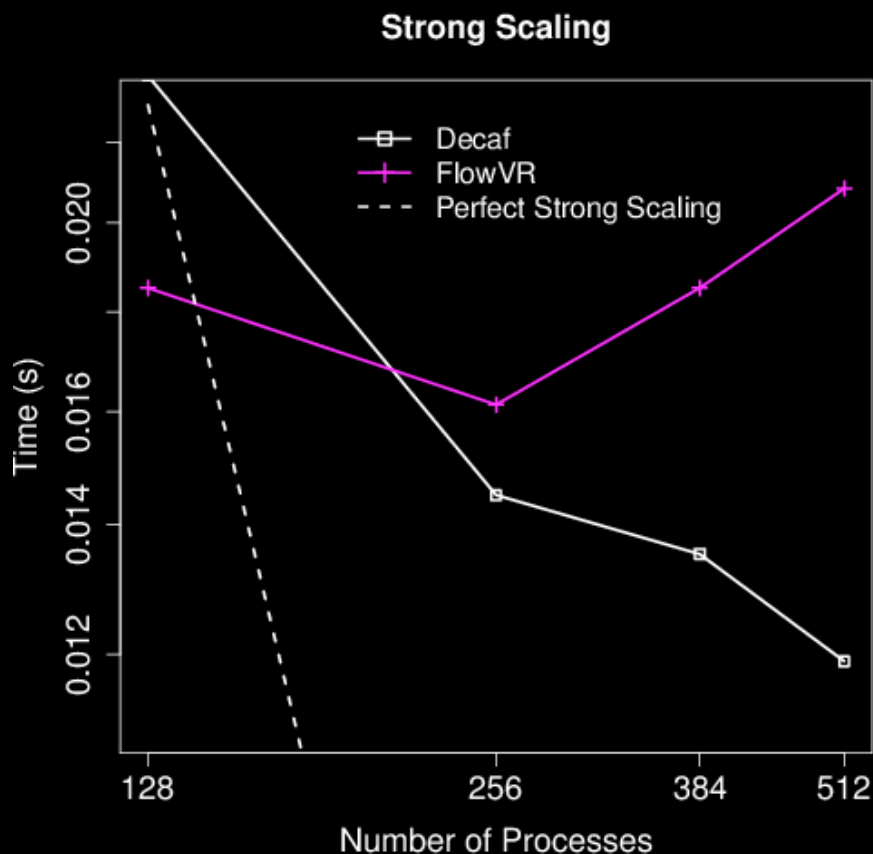


Decaf is a programming model and runtime for coupling HPC codes.

- Decoupled workflow links with configurable dataflow
- Data redistribution patterns
- Flow control
- Resilience

Three Examples in Greater Detail

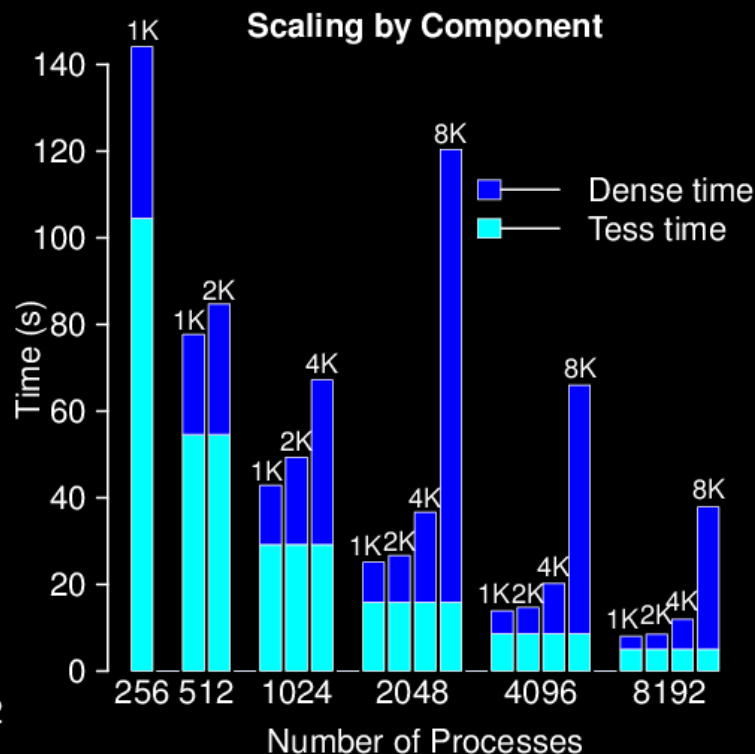
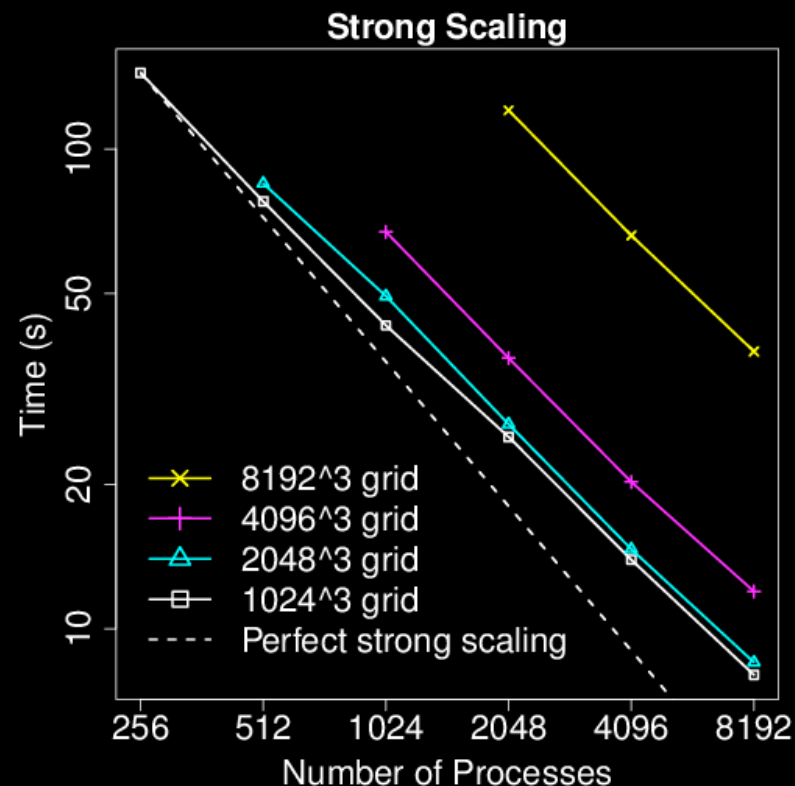
Data Redistribution in Molecular Dynamics



Three different redistributions are performed while computing an isosurface from an MD simulation of 54,000 lipids (2.1 M particles). [Dreher et al. 2014]

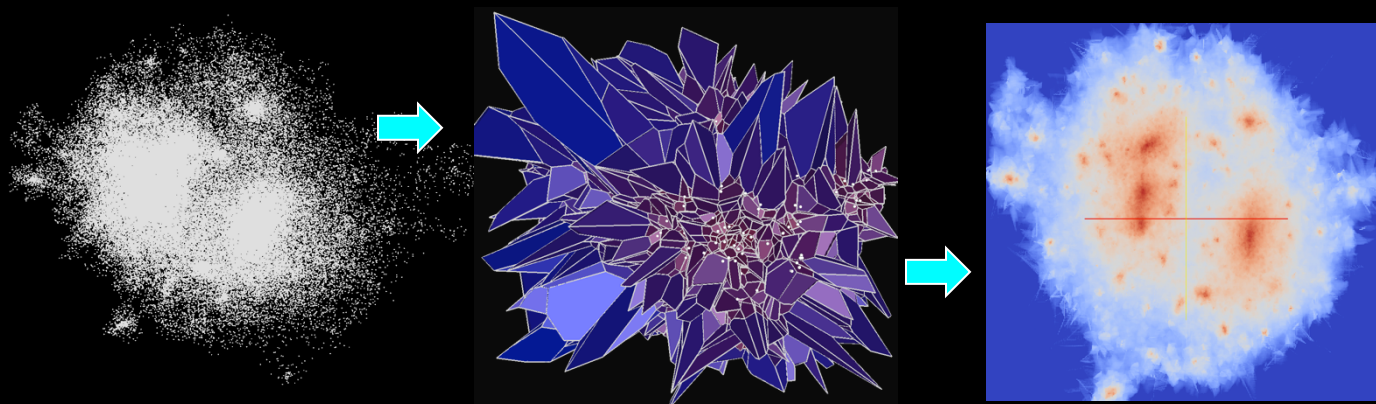
We applied the decaf redistribution library to the Gromacs molecular dynamics code in order to visualize isosurfaces from molecular density. Code complexity was reduced dramatically, while maintaining performance improved.

Density Estimation in Cosmology



Left: Strong scaling for 512³ synthetic particles various grid sizes.

Right: Scaling of individual tessellation and density estimation components.

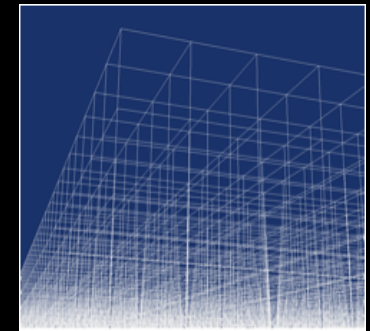
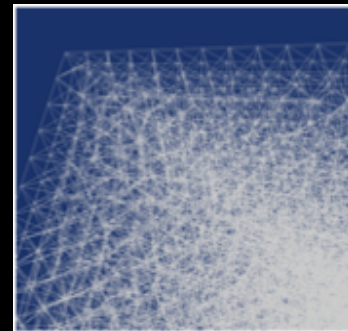
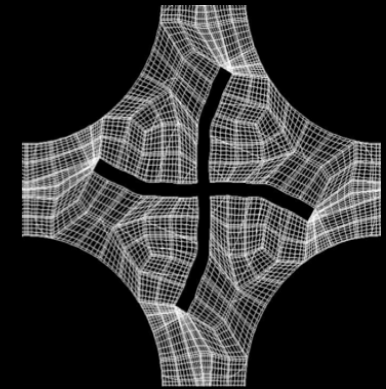
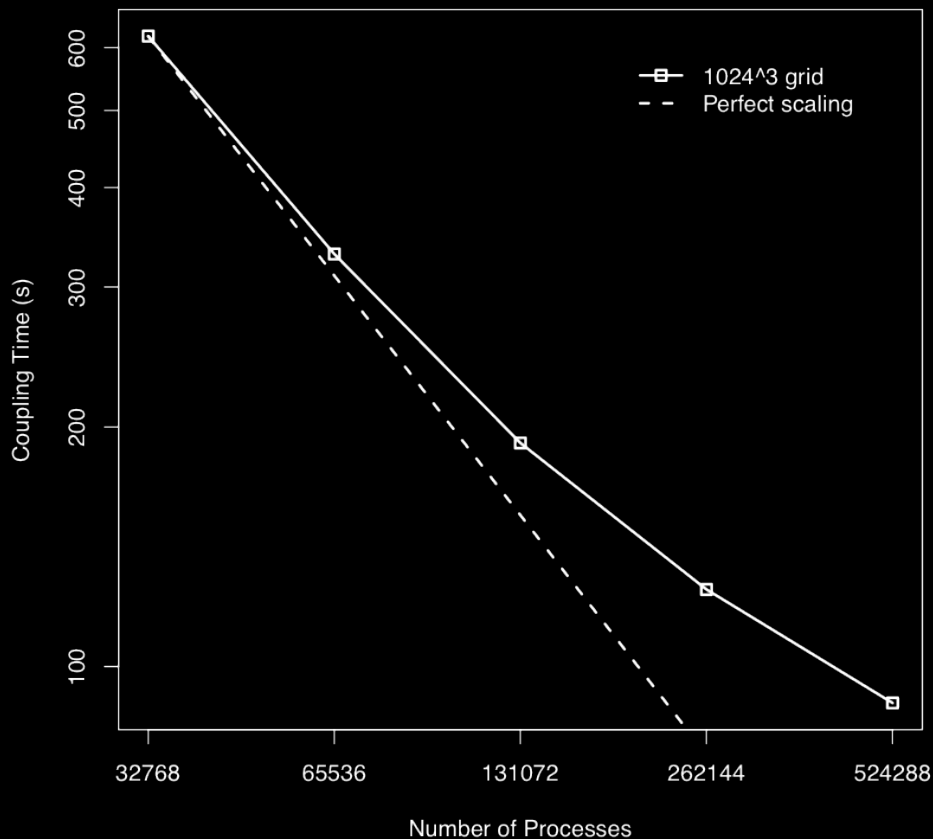


Density estimation: Tessellations as intermediate representations enable accurate regular grid density estimators.

Multiphysics Code Coupling in Nuclear Engineering

The cian proxy app of the CESAR codesign center emulate multiphysics coupling between neutronics and thermal hydraulics in nuclear reactor design.

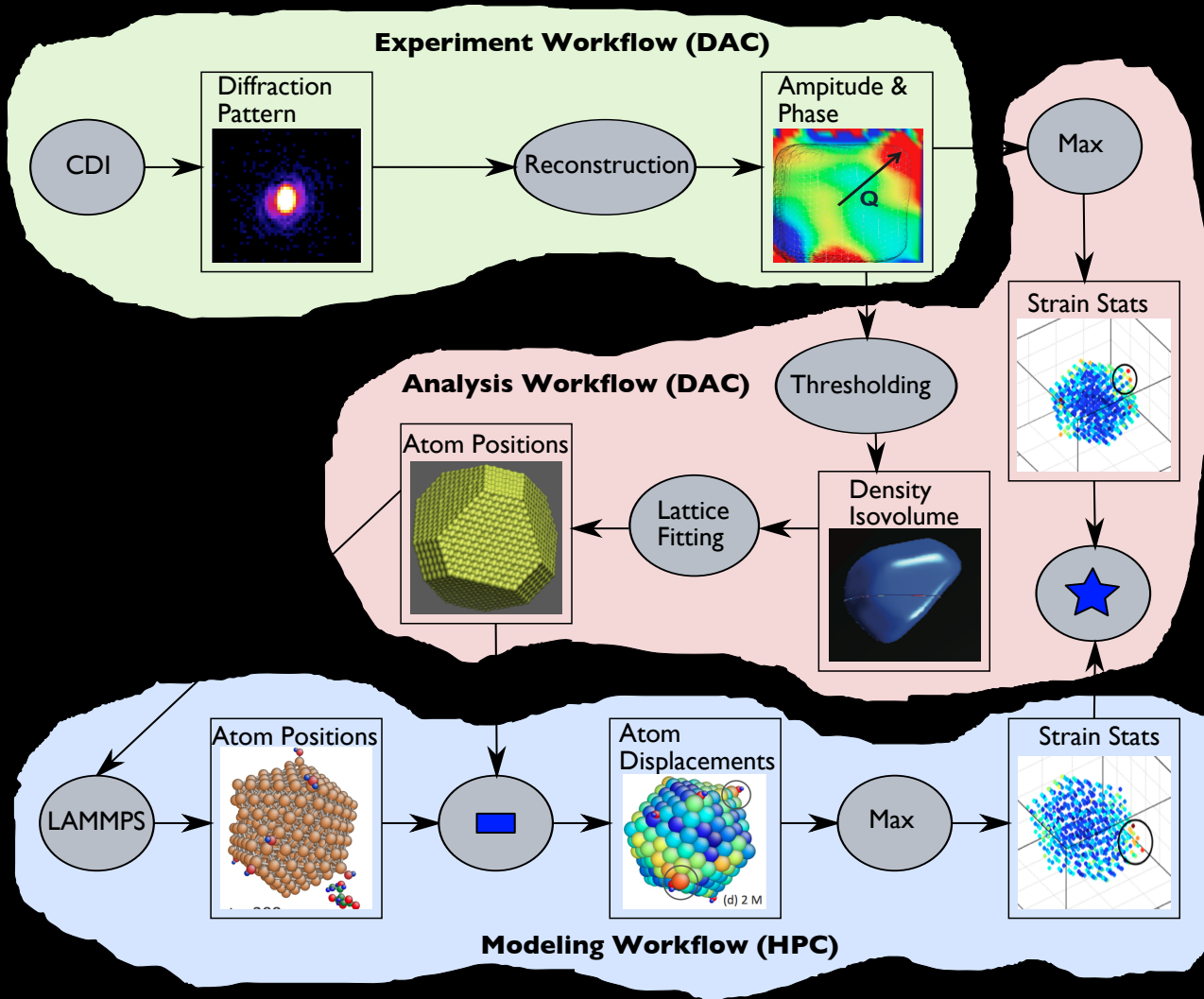
Strong Scaling



Coupling proxy app to transfer a solution from a 1024^3 tetrahedral mesh to a 1024^3 hexahedral mesh and back again at up to $\frac{1}{2}$ million blocks (MPI processes) and 43% strong scaling efficiency.

Advanced Topics

Combining Simulation and Experiment



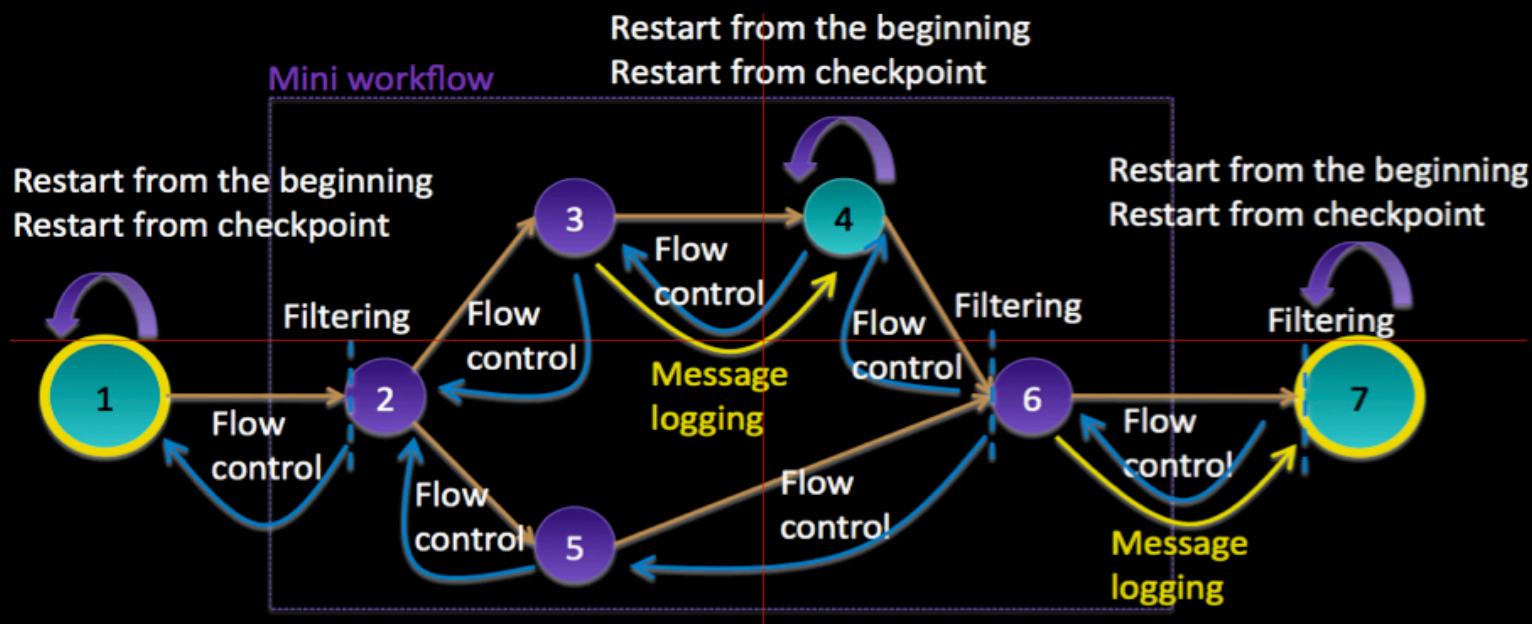
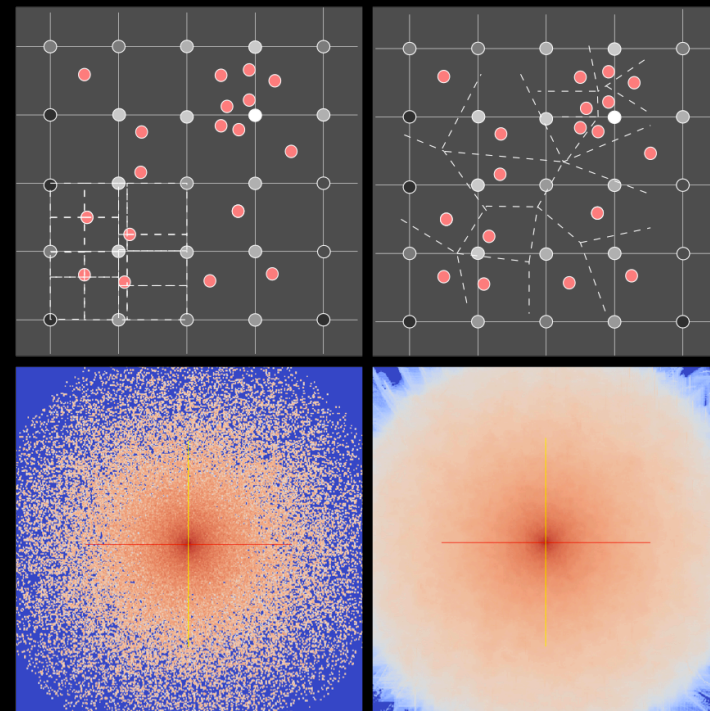
Science workflow for the comparison of a molecular dynamics simulation with a high-energy X-ray microscopy of the same material system includes three interrelated computational and experimental workflows.

CS problem: How to combine different (HPC and DAIC) WMSs?

Resilience to Faults

One of our resilience efforts attempts to detect silent data corruption by validating analysis tasks with an auxiliary method, usually less expensive and less accurate, but hopefully good enough to detect soft errors.

Another research topic is modeling the dataflow and optimally adding replication and roll back mechanisms to recover from hard (fail stop) errors and soft errors detected above.



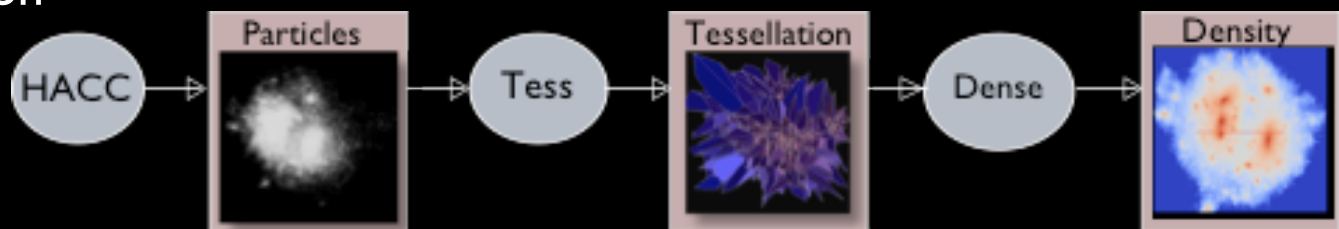
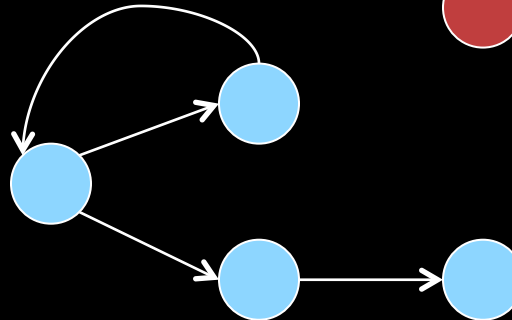
Mini Workflows

Recipe

- Take examples/tutorials from software libraries
- Add timing and other measurements
- Add parameters to make various configurations
- Add documentation, release as a benchmark suite

Decaf examples

- Linear 2 nodes
- Linear 3 nodes
- Cycle 4 nodes
- HACC
- LAMMPS
- Navier-Stokes CFD
- Redistribution



Wrap up

ModSim Challenges for In Situ Workflows

- **Data Movement**

- Data redistribution between pairs of tasks (semantic-preserving)
- Parallel data movement design patterns (direct, N:M, aggregate, buffer, pipeline, permute)

- **Resource selection, provisioning, scheduling (compute, network, storage)**

- Interfaces to scheduling systems, coordination of data transfers and task scheduling
- Future is to move from static to dynamic resource allocation

- **Validation**

- Infrastructure and application monitoring
- Understanding workflow behavior (modeling, anomaly detection and diagnosis)
- Correctness and performance expectations are fuzzy

- **Provenance capture**

- Fast storage and retrieval
- Analysis and mining

- **Productivity, portability**

- UI and programming models not a Modsim concern, or are they?

Further Reading

- Deelman, E., Peterka, T., et al.: The Future of Scientific Workflows. Report of the DOE NGNS/CS Scientific Workflows Workshop, 2016.
- Wozniak, J., Peterka, T., Armstrong, T., Dinan, J., Lusk, E., Wilde, M., Foster, I.: Dataflow Coordination of Data-Parallel Tasks via MPI 3.0. EuroMPI, 2013.
- Dorier, M., Dreher, M., Peterka, T., Wozniak, J., Antoniu, G., Raffin, B.: Lessons Learned from Building In Situ Coupling Frameworks. Proceedings of ISAV 2015.
- Peterka, T., Croubois, H., Li, N., Rangel, E., Cappello, F.: Self-Adaptive Density Estimation of Particle Data. SIAM SISC 2016.
- Dreher, M., Peterka, T.: Bredala: Semantic Data Redistribution for In Situ Applications. To appear in Proceedings of IEEE Cluster 2016, Taipei, Taiwan, 2016.
- Dorier, M., Antoniu, G., Cappello, F., Snir, M., Sisneros, R., Yildiz, O. Ibrahim, S., Peterka, T., Orf, L.: Damaris: Addressing Performance Variability in Data Management for Post-Petascale Simulations. To appear in ACM ToPC journal, 2016.

Acknowledgments

Facilities

Argonne Leadership Computing Facility (ALCF)
Oak Ridge National Center for Computational Sciences (NCCS)
National Energy Research Scientific Computing Center (NERSC)

Funding

DOE SDMAV Exascale Initiative
DOE SciDAC SDAV Institute

People

Decaf: Franck Cappello (ANL), Jay Lofstead (SNL)

Tom Peterka

tpeterka@mcs.anl.gov

<https://bitbucket.org/tpeterka1/decaf>

<http://www.mcs.anl.gov/~tpeterka>

Mathematics and Computer Science Division